-5- (WPAT)
AN - 96-398728/40
XRPX- N96-336009
TI - Document search method for e.g. newspaper report, patent specification - producing index corresp. to extended partial character string and adding single character in partial character string when index size is larger than standard index size
DC - T01
PA - (HITA) HITACHI LTD
PR - 95.01.12 95JP-019673
NUM - 1 patent(s) 1 country(s)
PN -- JP08194718 A 96.07.30 * (9640) 36p G06F-017/30
AP -- 95JP-019673 95.01.12
IC1 - G06F-017/30
AB - JP08194718 A

The method involves gathering the document data stored in a text database as the character code data. The position data containing the character position and the text specification data in a text data on a partial character string are extracted and stored as an index. The stored position data of the partial character string and the index are then managed and a character string table is updated. A predetermined partial character string is extracted from the search term and an applicable index is read with reference to the character string table.

The position data which has the same position relation as the partial character string in the search term is extracted. The size of the index corresp. to the partial character string produced from the text data is compared with the predetermined index size. The index corresp. to the extended partial character string is produced and a single character is added to the partial character string when the index size is larger than the standard index size.

ADVANTAGE - Minimises index capacity by lengthening character string even when large scale database is operated. Provides high speed searching for registration in large scale document database. (Dwg.2/41)

FN - WPG8JNS1.GIF

-1- (INSM)
AN - 6225516
ABN - C1999-05-6160D-013
TI - Integrating SQL databases with content-specific search engines.
AU - Dessloch S; Mattos N
ED - Jarke M; Carey M; Dittrich KR; Lockovsky F; Loucopoulos P; Jeusfeld MA
OS - IBM Database Technol. Inst., San Jose, CA, USA
SO - Proceedings of the Twenty-Third International Conference on Very Large Databases, pp. 528-537, Published: San Francisco, CA, USA, 1997, xvi+599 pp.
PU - Morgan Kaufmann Publishers
CP - USA
LA - English
DT - PA (Conference Paper)
NU - ISBN 1558604707
PY - 97
CONF- Proceedings of the Twenty-Third International Conference on Very Large Databases, Athens, Greece, 26-29 Aug. 1997
TC - PR (Practical); TM (Theoretical/Mathematical)
AB - In recent years, database research and product development activities have focused on support for non-traditional data types, such as text or multi-media documents. The paper describes an approach of coupling SQL databases and content-specific search engines, such as full-text retrieval engines, in an efficient manner. It is based on a query rewrite scheme that exploits so-called table functions, which are used to pass results from external search engines into the database engine. Using this approach the content-specific indexing mechanisms of search engines can be exploited without having to extend the database engine with new access methods, or having to break up the search engine to map its indexing scheme to database index structures. (28 Ref.)
IT - database indexing; full-text databases; query processing; relational databases; rewriting systems; search engines; SQL
ST - SQL databases; content-specific search engines; full-text retrieval engines; query rewrite scheme; table functions; external search engines; database engine; content-specific indexing mechanisms
CC - C6160D Relational databases;
C7250N Front end systems for online searching;
C4210L Formal languages and computational linguistics

-2- (INSM)
AN - 5471222
ABN - C9702-7250R-011
TI - Fuzzy full-text searches in OCR databases.
AU - Myka A; Guntzer U
ED - Adam NR; Bhargava BK; Halem M; Yesha Y
OS - Wilhelm-Schickard-Inst. fur Inf., Tubingen Univ., Germany
SO - Digital Libraries. Research and Technology Advances. ADL '95 Forum. Selected Papers, pp. 131-145, Published: Berlin, Germany,

-2- (INSM)
AN - 5471222
ABN - C9702-7250R-011
TI - Fuzzy full-text searches in OCR databases.
AU - Myka A; Guntzer U
ED - Adam NR; Bhargava BK; Halem M; Yesha Y
OS - Wilhelm-Schickard-Inst. fur Inf., Tubingen Univ., Germany
SO - Digital Libraries. Research and Technology Advances. ADL '95 Forum. Selected Papers, pp. 131-145, Published: Berlin, Germany, 1996, xiii+290 pp.
PU - Springer-Verlag
CP - Germany
LA - English
DT - PA (Conference Paper)
NU - ISBN 3540614109
PY - 96
CONF- Digital Libraries. Research and Technology Advances. ADL '95 Forum. Selected Papers, McLean, VA, USA, 15-17 May 1995, Sponsored by: NASA
TC - PR (Practical); XP (Experimental)
AB - We describe several methods of fuzzy full-text searching that can be used in OCR databases. The preliminary results of our tests show that it is possible to increase recall to a certain extent without losing too much precision. However, because almost every increase of recall has to be paid for by a decrease of precision, the user or administrator, respectively, of an information retrieval system has to decide individually on the importance of each of these values: if the risk of decreasing precision should be minimized, a slight increase of recall can be achieved by means of using character classes. If the highest possible recall (together with a reasonable precision) is necessary, employment of a confusion table may be the appropriate choice. Of course, in the case when significant loss of precision is introduced into the system by means of a fault-tolerant string matching algorithm, additional tools have to be provided to ensure that a user is not flooded with unnecessary information, e.g., a KWIC index could be used in order to provide for a preliminary overview of results. The tests are still extended at the moment. Extensions include the size of the analyzed full-text database as well as the addition of variations of some of the described methods, especially those based on linear scanning. However, such a test can never be complete: different algorithms can be evaluated as well as variations and combinations of the ones described here. Furthermore, OCR errors are, to a certain extent, dependent on the OCR device. Therefore, different results could possibly be achieved by means of testing data from different OCR devices. (0 Ref.)
IT - character sets; full-text databases; fuzzy logic; optical character recognition; query formulation; string matching
ST - fuzzy full-text searches; OCR databases; recall; precision; information retrieval system; character classes; confusion table; fault-tolerant string matching algorithm; KWIC index; linear scanning
CC - C7250R Information retrieval techniques

CPR - Copyright 1997, IE

-3- (INSM)
AN - 4990096
ABN - C9508-7250-008
TI - Efficient signature file methods for text retrieval.
AU - Dik Lun Lee; Young Man Kim; Gaurav Patel
OS - Dept. of Comput. & Inf. Sci., Ohio State Univ., Columbus, OH, USA
SO - IEEE Transactions on Knowledge and Data Engineering, vol.7, no.3, pp. 423-435, June 1995
CP - USA
LA - English
DT - J (Journal Paper)
NU - ISSN 1041-4347
PY - 95
TC - PR (Practical)
CPN - 1041-4347/95/ $04.00
AB - Signature files have been studied extensively, as an access method for textual databases. Many approaches have been proposed for searching signatures files efficiently. However, different methods make different assumptions and use different performance measures, making it difficult to compare their performance. In this paper, we study three basic methods proposed in the literature, namely, the indexed descriptor file, the two-level superimposed coding scheme, and the partitioned signature file approach. The contribution of this paper is two-fold. First, we present a uniform analytical performance model so that the methods can be compared fairly and consistently. The analysis shows that the two-level superimposed coding scheme, if stored in a transposed file, has the best performance. Second, we extend the two-level superimposed coding method into a multilevel superimposed coding method, we obtain the optimal number of levels for the multilevel method and show that for databases with reasonable size the optimal value is much larger than 2, which is assumed in the two-level method. The accuracy of the analytical formula is demonstrated by simulation. (21 Ref.)
IT - information retrieval
ST - signature file methods; text retrieval; access method; textual databases; performance measures; indexed descriptor file; two-level superimposed coding scheme; partitioned signature file approach; simulation
CC - C7250 Information storage and retrieval
CPR - Copyright 1995, IE

-4- (INSM)
AN - 2221231
ABN - C84018696
TI - Interactive information management system: concept and design.
AU - Goldstein CM
ED - Keren C; Perlmutter L
OS - Nat. Library of Medicine, Nat. Insts. of Health, Bethesda, MD, USA
SO - Application of Mini- and Micro-Computers in Information, Documentation and Libraries. Proceedings of the International